Resurrecting an Ancient Library: Detection and Decoding of Ink from 3D X-ray Scans of Carbonized Scrolls

[1]*Yu-Fan Ling* (凌宇帆) [2]*Chiou-Shann Fuh* (傅楸善)

[1]Graduate Institute of Bio-electronics and Bioinformatics
[2]Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
r11945060@ntu.edu.tw fuh@csie.ntu.edu.tw

ABSTRACT

We present and evaluate the approach employed in the Vesuvius Challenge - Ink Detection, a Kaggle competition aimed at resurrecting an ancient library from the ashes of a volcano. The competition tasked participants with detecting ink from 3D X-ray scans and deciphering the contents of thousands of scrolls belonging to a library in a Roman villa in Herculaneum, a town near Pompeii. These scrolls were carbonized and rendered fragile due to the Vesuvius eruption nearly 2000 years ago, making them impossible to open without damage. Fig. 1 shows a scroll that cannot be opened. Our approach involves utilizing advanced machine learning techniques, including convolutional neural networks and computer vision algorithms, to analyze the 3D X-ray scans and extract the ink traces. By accounting for the varying conditions of the scrolls, our model adapts to the challenges of carbonization, ink degradation, and fragmentation. We won a bronze medal in the competition for making a significant contribution to understanding the ancient library and unlocking previously unobtainable knowledge.

Fig. 1. One of the scrolls cannot be physically opened. [1]

Keywords: *Image segmentation, Kaggle, CVGIP 2023*

1. INTRODUCTION

This study addresses the challenge of recovering ink presence from 3D X-ray scans of detached fragments of ancient papyrus scrolls, a crucial subproblem in solving the Vesuvius Challenge. The primary goal is to detect and differentiate ink from papyrus using advanced image segmentation techniques [2]. The competition organizers provided a dataset consisting of 3D X-ray scans of ancient papyrus scroll fragments, binary masks indicating pixels with data, ink vs no-ink labels, infrared photos on which the binary masks are based, and run-length-encoded versions of the labels. This comprehensive dataset enables participants to develop and apply advanced image segmentation techniques to accurately detect ink presence and decipher the contents of these historically significant scrolls. We experiment with three distinct approaches: 2D, 2.5D, and 3D methods, employing U-Net [3] for the first two and a custom InkClassifier3DCNN from ink-id for the last one. By comparing the performance of these three approaches, we aim to identify the most suitable method for recovering ink from 3D X-ray scans of papyrus scrolls, contributing to the ongoing efforts to decipher and preserve ancient texts.

2. RELATED WORK

Virtual unwrapping [4, 5], a groundbreaking approach that combines heritage science and computational methods, utilizes non-invasive volumetric imaging and a multi-step computational process to delineate three-dimensional surfaces, unveil concealed text, and convert the findings into two-dimensional images. This innovative technique has showcased its adaptability to diverse challenges and materials through its successful application to artificially created manuscripts in the laboratory. These initial experiments have demonstrated the method's flexibility across various imaging techniques, substrates, inks, and manuscript formats such as scrolls and folded sheets. As the field progressed, it advanced to tackle more intricate materials like bamboo scrolls, metallic inks on parchment and rolled or folded papyri. These significant developments have paved the way for the successful recovery of texts from authentic heritage manuscripts, including those composed of metallic inks on parchment, paper, papyri, etched metal scrolls, and lead amulets.

A key study in this area, entitled "EduceLab-Scrolls: Verifiable Recovery of Text from Herculaneum Papyri using X-ray CT," was conducted by Parsons et al. at the University of Kentucky [6]. This paper delves into the use of X-ray Computed Tomography (CT) scans to recover text from the Herculaneum Papyri, providing valuable insights into the potential of X-ray imaging technologies for ancient text recovery.

Mocella et al. contributed to the field "Revealing letters in rolled Herculaneum papyri by X-ray phase-contrast imaging" [7]. This research demonstrated the effectiveness of phase-contrast imaging, an X-ray technique, in discerning letters in rolled Herculaneum papyri, thus showing this method's potential in restoring ancient texts.

The study "Virtual unrolling and deciphering of Herculaneum papyri by X-ray phase-contrast tomography" by Bukreeva et al. further underscored the application of X-ray phase-contrast tomography in virtually unrolling and deciphering the Herculaneum papyri, showcasing the possibilities of integrating virtual unrolling techniques with X-ray imaging [8].

In a different direction, the study "Unlocking history through the automated virtual unfolding of sealed documents imaged by X-ray microtomography" by Dambrogio et al. explored the potential of automated virtual unfolding [9]. Using X-ray microtomography, the authors achieved significant progress in unveiling historical documents that were initially inaccessible due to their sealed status.

While these studies have advanced the field, there is still room for further exploration. In particular, integrating machine learning segmentation techniques in the processing of 3D X-ray scans presents promising opportunities for improving the accuracy and efficiency of text recovery. Machine learning segmentation techniques, such as Convolutional Neural Networks (CNNs) and U-Net architectures, have shown great promise in other areas of image segmentation and could be applied to this field.

In the problem of image segmentation, the process of partitioning an image into multiple segments or sets of pixels has been extensively studied in computer vision, with applications ranging from medical imaging to object recognition. 2D and 3D image segmentation techniques have been developed and utilized for different applications [10].

In the context of 2D image segmentation, thresholding, edge-based, region-based, and clustering methods have been widely employed. For instance, Otsu's method [11] is a threshold-based technique that has been used for the binarization of grayscale images. Edge-based methods, like the Canny edge detector [12], detect discontinuities in brightness. Like the Watershed algorithm [13], region-based techniques separate images into regions based on pixel intensity. Clustering methods like K-means [14] clustering have also been used to partition an image into clusters based on similarity in pixel values.

Turning 3D image segmentation involves segmenting volumetric images [15, 16] into structures of interest. Techniques include thresholding, edge detection, region growing, and watershed. 3D image segmentation has significant applications in medical imaging, such as CT and MRI [17], where structures of interest must be isolated for further analysis.

In the specific case of detecting and decoding ink from 3D X-ray scans of carbonized scrolls, the task becomes even more challenging due to the scrolls' delicate nature and extreme fragility. Past work has involved applying machine learning techniques and advanced image processing algorithms to detect traces of ink and discern the underlying text.

Despite the advances in these methods, challenges persist in the accurate segmentation of ink from 3D X-ray scans of carbonized scrolls, particularly in dealing with the high level of noise, the complexity of the scroll structure, and the faintness of the ink traces.

3. DATA

This competition has a robust training dataset, a critical component of our research, assembled from four distinctive elements, each contributing uniquely to the overall richness of the dataset.

The first part is the surface volumes generated using advanced 3D X-ray scanning technology applied to the papyrus. This component comprises a set of 65 grayscale images. Fig. 2 is an example, each captured precisely and stored as uint16 data. These images thoroughly represent the papyrus's intricate structure, serving as a gateway to uncovering its hidden narratives and historical insights.

The second component is the ink labels, like Fig. 3, meticulously applied by hand to pinpoint the exact locations of the ink on the papyrus. Critical for their precision and dependability, these labels are a key reference within our dataset. They assist in identifying and examining various script patterns and embed textual content in the papyrus.

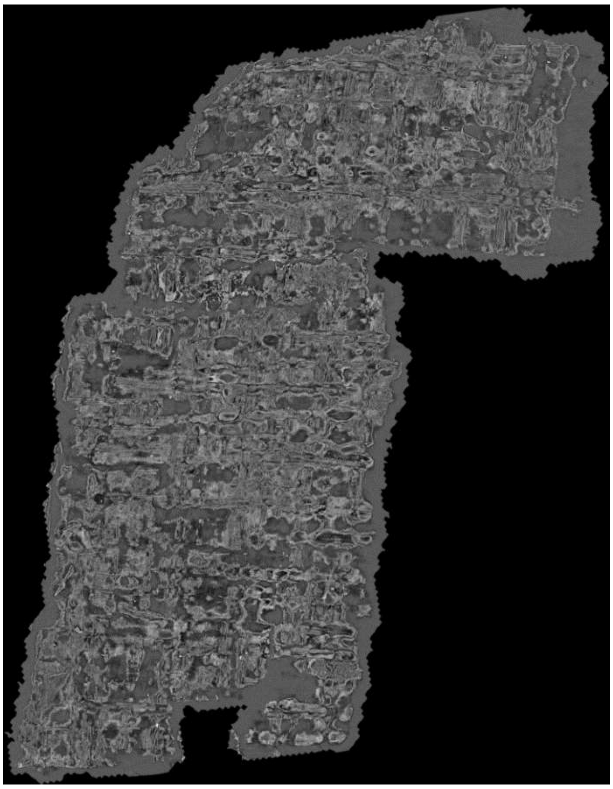


Fig. 2. A sample of the greyscale slice.

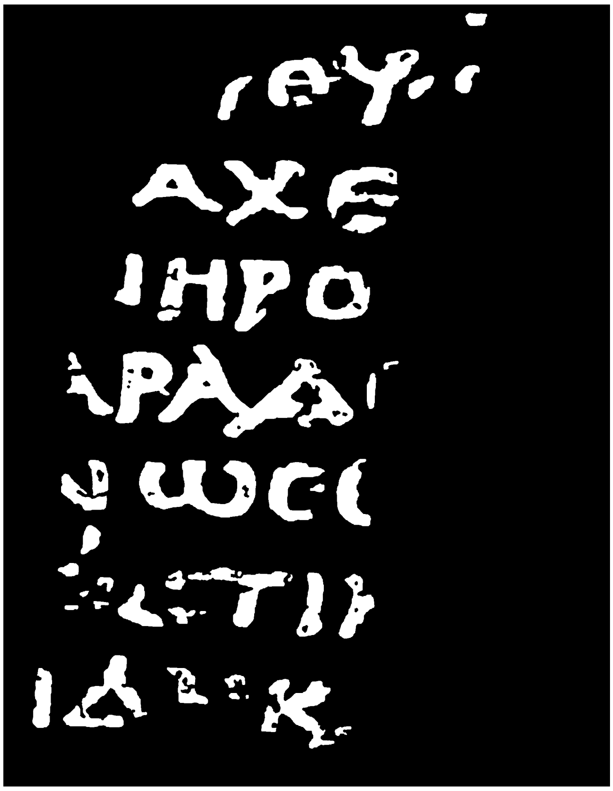


Fig. 3. A binary mask of the ink vs no-ink labels.

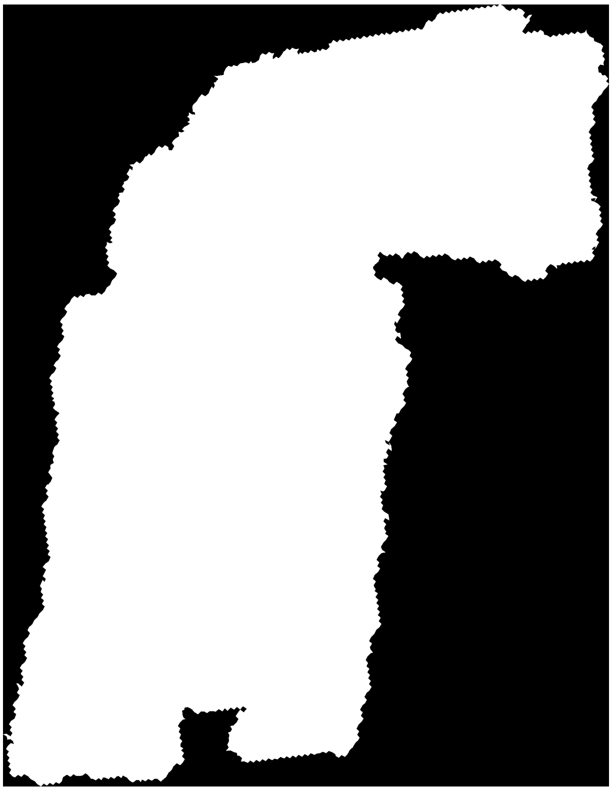


Fig. 4. A binary mask of which pixels contain data.

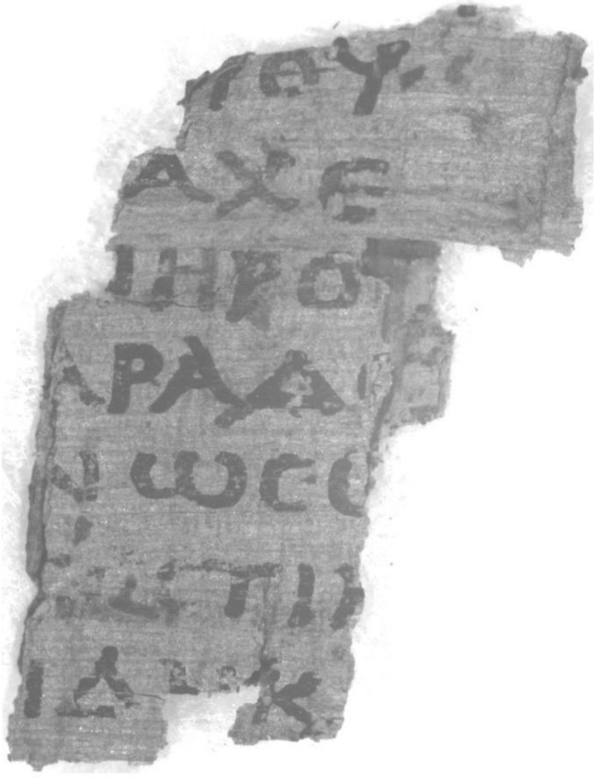


Fig. 5. The infrared photo on which the binary mask is based.

Thirdly, we introduce mark images to our dataset, as shown in Fig. 4. Given our scans' opaque and predominantly black backdrop, these images serve as visual markers, illuminating and demarcating the specific areas subjected to the scanning process.

Lastly, we incorporate infrared photos, forming the basis for our binary masks. These photos, shown in Fig. 5, provide a different perspective on the papyrus, revealing details that are not easily visible in other imaging modalities. They serve as an essential tool for generating binary masks, which help differentiate the areas of interest in the papyrus from the background.

In combination, these four elements create our comprehensive training dataset. This multifaceted resource significantly contributes to the study and understanding of papyrus.

4. SEGMENTATION METHODS

In ink detection tasks, since the train data are slices from the 3D x-ray surface volume, methods of different dimensions can be used to analyze and identify ink regions in a given image. The 2D methods focus on individual image slices, while the 2.5D methods [17] leverage a few adjacent slices to capture spatial information. The 3D methods, on the other hand, consider the entire volume to model the spatial relationships among all image slices. By comparing these techniques, we can choose a more suitable technique to capture the spatial context and improve the accuracy of ink detection in the dataset.

4.1. 2D SEGMENTATION

Our research introduces a 2D segmentation method based on the U-Net architecture for ink identification tasks. This method focuses on the U-Net structure and the use of Binary Cross-Entropy Loss (BCELoss) for training the model. This is the most intuitive approach, which is to process each slice from the 3D X-ray surface volume individually. We have implemented and trained the model on a comprehensive dataset, ensuring it can effectively segment and identify ink types in 2D images.

The U-Net architecture is a well-established convolutional neural network designed specifically for biomedical image segmentation tasks. It comprises an encoder-decoder structure with skip connections, enabling the model to capture high-level features and fine-grained details from the input images. The encoder consists of multiple convolutional and pooling layers, which help the model learn complex patterns and extract relevant features from the input data. On the other hand, the decoder consists of up-convolutional layers and skip connections that allow the model to reconstruct the segmented image with precise details.

4.2. 2.5D SEGMENTATION

We propose a 2.5D segmentation method that focuses on improving the performance of segmentation tasks by incorporating the U-Net architecture, using only 6 slices from the original data, applying slide inference, and adding rotation-based Test-Time Augmentation (TTA) [18].

In this approach, we also use the U-Net model, and by using only 6 slices in the middle, we aim to reduce the computational burden while maintaining sufficient information for accurate segmentation. This is achieved by selecting the central slices from the 3D input data and converting them into a 2.5D representation.

Slide inference is employed to handle large-sized images, divided into smaller overlapping tiles processed individually by the model. The final segmentation mask is then reconstructed by merging the predictions of these tiles while accounting for the overlapping regions. This approach allows for the effective handling of large images without incurring memory issues.

Additionally, we introduce rotation-based TTA to enhance the model's robustness to various orientations. During inference, the input image is rotated by multiple angles, and the model's predictions for each rotated image are averaged to obtain the final output. This approach has improved the model's generalization capabilities, mitigating the potential overfitting to a specific orientation.

4.3. 3D SEGMENTATION

We propose a 3D segmentation method using a 3D Convolutional Neural Network (3D CNN) for the ink identification task inspired by the InkClassifier3D CNN model from the ink-id repository. The primary goal of this method is to enhance the baseline score of 0.48 achieved by the initial model. We have implemented and adapted the model using a Kaggle notebook, ensuring that it remains within the computational limits of the platform.

The proposed method takes an input sub-volume and generates a prediction for a single pixel within the 3D volume. We have devised a sparse prediction technique to achieve efficient prediction for the test fragments while adhering to the computational constraints of a Kaggle notebook. In this approach, we predict every 19th pixel and extrapolate the data to the surrounding areas by replicating the predicted pixel values. This strategy allows for faster computation without significantly compromising the model's performance. We have demonstrated the effectiveness of this technique using validation data withheld from the model during training.

The core of our method lies in the 3D CNN [19] architecture, specifically designed for the ink identification task. The 3D CNN model captures the spatial relationships among the input voxels, allowing it to better differentiate between various ink types in the volume. The model's architecture consists of multiple convolutional, pooling, and fully connected layers, enabling it to learn complex patterns and features from the input data. The use of 3D CNNs in our segmentation method demonstrates its potential for improving the state-of-the-art ink identification and segmentation tasks, outperforming the baseline score achieved by the initial model.

4.4. 2.5D + 3D SEGMENTATION

We propose a hybrid segmentation method that combines 2.5D and 3D Convolutional Neural Networks (CNNs) for the ink identification task. This method is designed to leverage the strengths of 2.5D and 3D approaches, balancing computational efficiency and the ability to capture spatial relationships in the data.

The process begins with a 2.5D preprocessing step, where the 3D images are sliced into multiple 2D images. This approach allows us to handle the 3D data as a series of 2D slices, reducing the computational complexity. These 2D slices are fed into a 3D ResNet [20] model for feature extraction. The 3D ResNet model can capture spatial relationships between the slices, thus incorporating 3D information into the feature maps.

The feature maps are then passed through a decoder to generate the final segmentation masks. The masks are post-processed using a denoising function and thresholded to create binary masks. The final step involves converting the binary masks into run-length encoding (RLE) format for submission.

This hybrid 2.5D + 3D segmentation method effectively combines the benefits of 2.5D and 3D methods, allowing for accurate segmentation while maintaining computational efficiency. It demonstrates the potential of using a combination of 2.5D preprocessing and 3D CNNs for ink identification and segmentation tasks, providing an efficient approach to addressing our research objective.

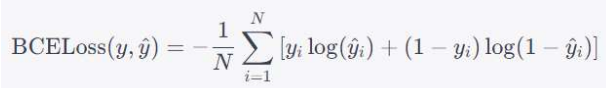
5. METHODS

To improve our score on Kaggle, after determining the segmentation method, we also adopted many training techniques. Some of these had a limited impact on the final improvement, while others could have a more substantial effect.

5.1. LOSS FUNCTION

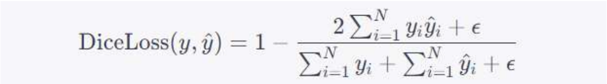
We use the following Combined Loss Function to achieve the best results:

BCELoss (Binary Cross Entropy Loss):



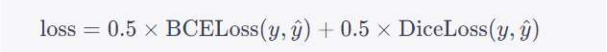
$$\mathrm{BCELoss}(y,\hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right]$$

DiceLoss:



$$\mathrm{DiceLoss}(y,\hat{y}) = 1 - \frac{2\sum_{i=1}^{N} y_i\hat{y}_i + \epsilon}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N}\hat{y}_i + \epsilon}$$

Combined Loss Function:



$$\mathrm{loss} = 0.5 \times \mathrm{BCELoss}(y,\hat{y}) + 0.5 \times \mathrm{DiceLoss}(y,\hat{y})$$

Using a combination of Binary Cross Entropy Loss (BCELoss) and Dice Loss in a 0.5:0.5 ratio can bring several advantages, especially in tasks like image segmentation:

Balance of different objectives: BCELoss measures the error between the predicted probabilities and accurate classifications, pushing the model to classify each pixel correctly. On the other hand, Dice Loss evaluates the overlap between the predicted and actual regions, encouraging the model to produce predictions that match the true shape as closely as possible. By combining these two, you ensure that both aspects are being optimized.

Mitigation of class imbalance: In many image segmentation tasks, there's often a significant class imbalance, where most pixels belong to the background class. BCELoss alone might lead the model to predict everything as the majority class. Dice Loss can help to alleviate this issue because it cares about the overlap of regions, which cannot be improved by predicting everything as the background.

Stability of training: Balancing two different loss functions might also make the training process more stable and robust.

Better generalization: The combined loss could lead to better generalization on the test set, as it's not focusing on a single type of error but rather a more comprehensive performance measure.

5.2. DENOISING

For enhancing the ink detection performance, denoising plays a pivotal role, exploiting the properties of the ink distribution, which are two-fold: sparsity and continuity. Sparsity suggests that most of the regions on the papyrus will not have ink. At the same time, continuity implies that the presence of ink in one pixel increases the likelihood of its presence in the neighboring pixel.

Our research used L1/Hessian denoising [21, 22]. L1/Hessian denoising reduces noise in Magnetic Resonance Images (MRI) based on compressive sensing with L1 and Hessian regularizations. Compressive sensing is a technique that exploits the sparsity of an image to reconstruct it from fewer measurements than traditional methods. L1 regularization is a term that penalizes the non-zero elements of an image to make it sparser and reduce noise. Hessian regularization is a term that uses the second-order derivatives of an image to preserve some details from being over-smoothed. The Hessian matrix is a matrix of partial derivatives that describes the local curvature of a function. The Hessian matrix's eigenvalues can indicate an image's edge information. L1/Hessian denoising combines these two regularizations to balance sparsity and smoothness in MRI denoising.

5.3. MODEL ARCHITECTURE

To improve the performance of the competition, we also tried many different deep-learning models, such as ResNet34d [23] and SE-ResNeXt-150. We experimented with the complete model architecture and modified these mature models to fit our dataset better. The experimental results are shown in the table below. ResNet34d and SE-ResNeXt-150 is the basic variant of ResNet, and the experimental results are not very good.

PVT v2 refers to the second version of the Pyramid Vision Transformer(PVT) model[24]. It achieves better training performance on datasets may due to several reasons. Firstly, it reduces the computational complexity of the attention layer from quadratic to linear, making the model more efficient and scalable. This improvement allows for faster training and inference time. Secondly, PVT v2 incorporates overlapping patch embedding and convolutional feed-forward networks. This design choice enables the model to capture more local and global information from the images, enhancing the feature representation. The overlapping patches help capture fine-grained details, while the convolutional feed-forward networks effectively refine and aggregate these features.

Additionally, PVT v2 benefits from pre-training on the ImageNet dataset. ImageNet pre-training provides a strong initialization for the model parameters, allowing the model to start with a good representation of visual features. This initialization helps PVT v2 converge faster and better generalize various downstream tasks.

The combined training of ResNet34d as an encoder and U-Net as a decoder works best, benefiting from the powerful feature representation of ResNet34d, the multi-scale feature fusion and symmetric design of U-Net, and the potential of pre-training and transfer learning. Together, these factors help improve the model's training performance in segmentation tasks.

Encoder	**Decoder**	**Score↑**
ResNet34d	None	0.54
SE-ResNeXt-150	None	0.53
PVT_V2	None	**0.58**
ResNet34d	U-Net	**0.58**

In subsequent experiments, we also found that adding a pooling layer behind the model can improve the final training results. The experimental results are shown in the table below.

Encoder	**Decoder**	**Score↑**
PVT_V2+pooled	None	0.60
PVT_V2+pooled	Deformer	0.61
ResNet34d+pooled	U-Net	**0.62**

After adding the pooling layer, the results of each method have been greatly improved. This may be because the pooling layer helps to reduce the spatial dimension of the features, capturing more abstract and high-level information. The pooling layer increases the receptive field by downsampling the feature maps, allowing the model to capture more considerable contextual information. This enhanced feature representation enables the model to better distinguish and classify complex patterns in the input data, improving performance in various tasks. Additionally, the pooling layer can help reduce the model's computational complexity and memory usage, making it more efficient during training and inference.

We also tried to use the deformer as a decoder for the architecture in the PVT V2 paper [25], but the experimental results were not very good. In the end, ResNet34d added a pooling layer as an encoder, and U-Net as a decoder achieved the best results.

6. RESULT

We did many experiments to get the final score, but each segmentation method's training methods and parameters are inconsistent in practice. We can only do our best to achieve the best score for each segmentation method for comparison. The final Kaggle scores are shown in the table below.

Segmentation Method	**Pubic Score↑**	**Private Score↑**
2D	0.367157	0.372334
2.5D	0.534766	0.446494
3D	0.482412	0.468229
2.5D + 3D	**0.625686**	**0.572173**

The results of our experiments show that the 2.5D + 3D segmentation method outperforms the other methods, achieving a public score of 0.625686 and a private score of 0.572173 on the Kaggle platform, and finally ranked 114/1,289 and won the bronze medal in the competition.

This demonstrates the potential of this hybrid approach for ink identification and segmentation tasks and suggests that it may be a promising direction for future research in this area.

7. FUTURE WORK

In this project, we refer to many state-of-the-art segmentation methods, some of which ideas have been applied in research, but some have not been satisfactory. These latest ideas should serve as some guidance for our future work. In future work, we will continue to study these state-of-the-art segmentation methods:

1. SegFormer [26]: We aim to integrate the SegFormer architecture, which combines the benefits of transformers and semantic segmentation, to improve the accuracy and efficiency of our ink detection model.

2. MaskFormer [27]: Another promising direction is incorporating MaskFormer, demonstrating that per-pixel classification can be improved using instance-wise masks, leading to more accurate segmentation results.

3. Diffusion Models [28]: Lastly, we plan to investigate the application of diffusion models in ambiguous medical image segmentation, which could help us better handle uncertainty and improve the robustness of our ink detection approach in challenging scenarios.

8. CONCLUSION

In this study, we have explored and implemented various segmentation methods for the ink identification task, including 2D, 2.5D, 3D, and a hybrid 2.5D + 3D approach. Each method has strengths and limitations, and its performance varies depending on the task's complexity and the data's nature.

The 2D segmentation method, while computationally efficient, needs the ability to capture the spatial relationships in the 3D data, which can limit its performance for tasks that require an understanding of 3D structures. The 3D segmentation method, on the other hand, can capture these spatial relationships but is computationally intensive, making it less suitable for large-scale or real-time applications.

The 2.5D segmentation method balances these two extremes by treating the 3D data as a series of 2D slices. This approach reduces the computational complexity while incorporating some 3D information into the model. However, it may only partially capture the 3D spatial relationships in the data.

Our proposed hybrid 2.5D + 3D segmentation method combines the strengths of both 2.5D and 3D approaches. It starts with a 2.5D preprocessing step to reduce computational complexity, then uses a 3D ResNet model to capture the spatial relationships between the slices. The resulting feature maps are decoded to generate the final segmentation masks, which are post-processed and thresholded to create binary masks.

In future work, we will delve deeper into state-of-the-art segmentation models and apply them to ink recognition tasks. It is believed that these studies will help us improve the model's performance and bring us closer to our goal of fully decoding the ancient papyrus text.

REFERENCES

[1] VisCenter. (2019). [Screenshot]. *EDUCE: Imaging the Herculaneum Scrolls.* YouTube. https://youtu.be/PpNq2cFotyY.

[2] Ronneberger O, Fischer P, Brox T. *U-net: Convolutional networks for biomedical image segmentation*[C]//Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015: 234-241.

[3] Sen P C, Hajra M, Ghosh M. *Supervised classification algorithms in machine learning: A survey and review*[C]//Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018. Springer Singapore, 2020: 99-111.

[4] Seales W B, Lin Y. *Digital restoration using volumetric scanning*[C]//Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries. 2004: 117-124.

[5] W.B. Seales. *Reading the invisible library: A retrospective. In Carl Brune and Caroline Foutch, editors, Modern Alchemy: New Technology for Museum Collections.Gilcrease Museum*, 2017.

[6] Parsons S, Parker C S, Chapman C, et al. *EduceLab-Scrolls: Verifiable Recovery of Text from Herculaneum Papyri using X-ray CT*[J]. arXiv preprint arXiv:2304.02084, 2023.

[7] Mocella V, Brun E, Ferrero C, et al. *Revealing letters in rolled Herculaneum papyri by X-ray phase-contrast imaging*[J]. Nature communications, 2015, 6(1): 5895.

[8] Bukreeva I, Mittone A, Bravin A, et al. *Virtual unrolling and deciphering of Herculaneum papyri by X-ray phase-contrast tomography*[J]. Scientific reports, 2016, 6(1): 27227.

[9] Dambrogio J, Ghassaei A, Smith D S, et al. *Unlocking history through automated virtual unfolding of sealed documents imaged by X-ray microtomography*[J]. Nature communications, 2021, 12(1): 1184.

[10] Shirly S, *Ramesh K. Review on 2D and 3D MRI image segmentation techniques*[J]. Current Medical Imaging, 2019, 15(2): 150-160.

[11] Bangare S L, Dubal A, Bangare P S, et al. *Reviewing Otsu's method for image thresholding*[J]. International Journal of Applied Engineering Research, 2015, 10(9): 21777-21783.

[12] Ding L, Goshtasby A. *On the Canny edge detector[J]. Pattern recognition*, 2001, 34(3): 721-725.

[13] Kornilov A S, Safonov I V. *An overview of watershed algorithm implementations in open source libraries*[J]. Journal of Imaging, 2018, 4(10): 123.

[14] Hamerly G, Elkan C. *Learning the k in k-means*[J]. Advances in neural information processing systems, 2003, 16.

[15] Li K, Wu X, Chen D Z, et al. *Optimal surface segmentation in volumetric images-a graph-theoretic approach*[J]. IEEE transactions on pattern analysis and machine intelligence, 2005, 28(1): 119-134.

[16] Mehra K, Soliman H, Sahoo S R. *Data Augmentation using Feature Generation for Volumetric Medical Images*[J]. arXiv preprint arXiv:2209.14097, 2022.

[17] Clarke L P, Velthuizen R P, Camacho M A, et al. *MRI segmentation: methods and applications*[J]. Magnetic resonance imaging, 1995, 13(3): 343-368.

[18] Han L, Chen Y, Li J, et al. *Liver segmentation with 2.5 D perpendicular UNets*[J]. Computers & Electrical Engineering, 2021, 91: 107118.

[19] Wang G, Li W, Aertsen M, et al. *Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks*[J]. Neurocomputing, 2019, 338: 34-45.

[20] Alakwaa W, Nassef M, Badr A. *Lung cancer detection and classification with 3D convolutional neural network (3D-CNN)*[J]. International Journal of Advanced Computer Science and Applications, 2017, 8(8).

[21] Qiu Z, Yao T, Mei T. *Learning spatio-temporal representation with pseudo-3d residual networks*[C]//proceedings of the IEEE International Conference on Computer Vision. 2017: 5533-5541.

[22] Micchelli C A, Shen L, Xu Y, et al. *Proximity algorithms for the L1/TV image denoising model*[J]. Advances in Computational Mathematics, 2013, 38: 401-426.

[23] Deng L, Zhu H, Yang Z, et al. *Hessian matrix-based fourth-order anisotropic diffusion filter for image denoising*[J]. Optics & Laser Technology, 2019, 110: 184-190.

[24] He K, Zhang X, Ren S, et al. *Deep residual learning for image recognition*[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[25] Wang W, Xie E, Li X, et al. *Pyramid vision transformer: A versatile backbone for dense prediction without convolutions*[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 568-578.

[26] Wang W, Xie E, Li X, et al. *Pvt v2: Improved baselines with pyramid vision transformer*[J]. Computational Visual Media, 2022, 8(3): 415-424.

[27] Xie E, Wang W, Yu Z, et al. *SegFormer: Simple and efficient design for semantic segmentation with transformers*[J]. Advances in Neural Information Processing Systems, 2021, 34: 12077-12090.

[28] Cheng B, Schwing A, Kirillov A. *Per-pixel classification is not all you need for semantic segmentation*[J]. Advances in Neural Information Processing Systems, 2021, 34: 17864-17875.

[29] Rahman A, Valanarasu J M J, Hacihaliloglu I, et al. *Ambiguous Medical Image Segmentation using Diffusion Models*[J]. arXiv preprint arXiv:2304.04745, 2023.